Heavy Traffic Analysis of a Production-Distribution System

Elena V. Krichagina
Institute of Control Sciences
Moscow, Russia
Lawrence M. Wein

Heavy Traffic Analysis of a Production-Distribution System

Elena V. Krichagina
Institute of Control Sciences
Moscow, Russia
Lawrence M. Wein

HEAVY TRAFFIC ANALYSIS OF A PRODUCTION-DISTRIBUTION SYSTEM

Elena V. Krichagina

Institute of Control Sciences

Moscow, Russia

and

Lawrence M. Wein

Sloan School of Management, M.I.T.

Cambridge, Massachusetts

June 13, 1994

Abstract

We consider an optimization problem for a multiclass queueing model of a decentralized production-distribution system: A single server can produce a variety of products, and completed units pass through an infinite server delay node on their way to a warehouse inventory. Each product has its own warehouse, demand for each product is satisfied by its associated inventory, and unsatisfied demand is backordered. Each warehouse operates independently under its own base stock policy: The initial inventory is set at a base stock level, and each unit of demand triggers an order with the server. The server processes the orders in a first-come first-served manner, and the optimization problem is to find the base stock level for each product to minimize the expected average inventory holding and backordering cost. A heavy traffic approximation of this problem yields an inventory process that possesses both diffusion and Gaussian components. Computational results are performed to assess the accuracy of our analysis.

1 Introduction

The travel time between nodes, or queues, in a queueing network model is typically assumed to be instantaneous. However, customers in most real queueing systems incur delays when traveling from one node to another. When these delays are essentially independent of the congestion in the network, as is often the case, the delays are appropriately modeled as infinite server nodes within the queueing network. Examples include the delay for a vehicle to travel between traffic intersections, a job to move between workstations in a factory, and data to be transmitted over a dedicated low speed transmission line. An infinite server node may also be used as an aggregate representation of a large number of non-bottleneck (that is, low utilization) workcenters in a large factory, such as a wafer fabrication facility. The heavy traffic approach developed by Harrison (1988) has been used in recent years (see, for example, Wein 1992a, Kelly and Laws 1993 and references therein) to control queueing networks that possess no infinite server nodes. The aim of this paper is to make a first step towards generalizing this approach to queueing networks that contain infinite server nodes.

Towards this end, we focus on a specific problem that is simple but well motivated: An idealized queueing model of a multi-product production-distribution system that consists of a single production facility and K warehouses. The facility produces K different products, and a completed unit of a given product is transported to its respective warehouse; hence, all inventories are held at the warehouses. These inventories service the actual customer demand, and unsatisfied demand is backordered. In our queueing model, the production facility is modeled as a single server, and each product has its own general service time distribution and renewal demand process. The delay incurred by a unit to travel from the facility to a warehouse is a random variable. As alluded to above, one could also interpret the model as a production-inventory system with a single bottleneck workstation, where the delay represents the aggregate sojourn time through the workstations downstream of the bottleneck.

The performance of such a system relies heavily upon the production control policy that is employed. We derive a control policy for a *decentralized system*: Each warehouse has an

2

initial *base stock* level of inventory, and whenever a demand for a unit of product i occurs, the inventory at warehouse i is depleted and the warehouse places an order with the production facility. The facility serves orders in a first-come first-served (FCFS) fashion and is idle if no orders are in its queue. The optimization problem for the decentralized problem is to derive a base stock level for each warehouse to minimize the average cost of holding and backordering inventory.

Since the optimization problem appears to be difficult to analyze without making additional distributional assumptions, we resort to a heavy traffic analysis. This approprimation procedure requires the server to be busy the great majority of time to meet average demand, and assumes that the travel delays are lengthy. Under these assumptions, the inventory process for each product, appropriately normalized, has a diffusion and a Gaussian component. Although we have been unable to derive the stationary distribution of this inventory process, our analysis reveals the structure of the limiting behavior of the production-distribution system. Under additional assumptions, we employ our analysis to numerically derive base stock levels, and find that they are very close in value to the optimal levels obtained via simulation.

We initially planned to also analyze a *centralized system*, where a controller has global information about the inventories of each product, and decides in a dynamic fashion which product, if any, to produce next. Comparison of these two systems would allow us to assess the value of centralized information in production-distribution systems; see Zheng and Zipkin (1990) and Wein (1992b) for similar comparisons for production-inventory systems, which are essentially production-distribution systems with no travel delays. During the course of our analysis, we developed a variant of a myopic policy proposed by Peña and Zipkin (1992). Although numerical results in Peña and Zipkin, Veatch and Wein (1993) and Krichagina et al. (1992) show that this policy performs well in a production-inventory setting, we found that our variant of the policy did not perform well for the production-distribution system; in fact, the system did not stabilize under this policy in our simulation study. It appears that the long travel delays lead to a system that is not easily controllable. Consequently,

this paper does not contain an analysis of the centralized problem.

In summary, control policies for queueing networks with infinite server delay nodes are very difficult to analyze; if the delays are not exponential, then the cardinality of the state space can become unwieldy. Unfortunately, our analysis suggests that these types of problems will not yield easily to heavy traffic analysis.

The remainder of this paper is organized as follows. The queueing control problem for the decentralized model is defined in Section 2, and a sequence of systems approaching heavy traffic is introduced in Section 3. An approximating fluid control problem and an approximating diffusion-Gaussian control problem are formulated in Section 4; the fluid analysis is a prerequisite to the more elaborate heavy traffic approximation, and allows the controls in the heavy traffic model to be expressed as deviations from the optimal fluid controls. Further analysis of the model is undertaken in Section 5, along with a computational study that assesses the accuracy of our results.

2   The Problem Formulation

The decentralized production-distribution system pictured in Figure 1 consists of a single machine producing K products. Units of product $i = 1, \ldots, K$ have a general service time distribution with service rate μ_i and squared coefficient of variation ν_{ai}^2. We let $A_i(t) = \max\{j : \xi_1^i + \ldots + \xi_j^i \le t\}$ denote the renewal process associated with the iid service times $\{\xi_j^i\}_{j \ge 1}$. Completed units incur a random delay before entering the finished good inventory at the appropriate warehouse. Let $\{\eta_j^i\}_{j \ge 1}$ be the sequence of iid delays for product i, and define the continuous distribution function $F_i^0(t) = P(\eta_1^i \le t)$. We denote the mean of this distribution by λ_i^{-1}. Each product has an independent renewal process $D_i(t) = \max\{j : \zeta_1^i + \ldots + \zeta_j^i \le t\}$, which is the number of product i units demanded up to time t. We let d_i denote the demand rate and ν_{di}^2 be the square coefficient of variation of the interdemand times.

We assume that warehouses operate independently. Each warehouse initially has C_i units of inventory, and each unit of demand for product i simultaneously triggers two events:

Product i's inventory is depleted and a production order is placed with the machine. The machine serves orders in a FCFS fashion and is idle if no orders are in the queue. If we let $T_i(t)$ represent the total time in $[0, t]$ devoted to the production of product i, then $A_i(T_i(t))$ is the total number of product i units produced up to time t. Denote by $Q_i(t)$ the number of product i orders in queue or in service at the machine at time t. Let $Z_i(t)$ be the number of units in product i's delay node. Then

$$Q_i(t) = D_i(t) - A_i(T_i(t)) \tag{2.1}$$

and

$$Z_i(t) = \sum_{j=1}^{A_i(T_i(t))} I(\eta_j^i + \tau_j^i > t) , \tag{2.2}$$

where τ_j^i is the time of the jth jump of the process $A_i(T_i(\cdot))$; it corresponds to the epoch when the jth unit of product i completes production.
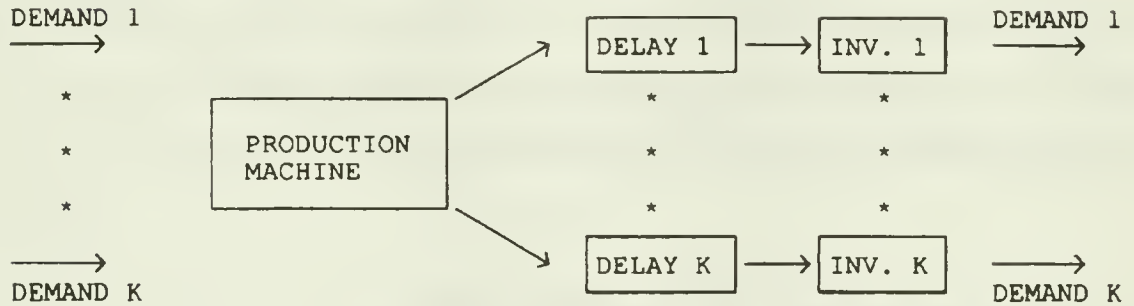

Figure 1. The production-distribution system.

Let $X_i(t)$ denote the number of units in product i's inventory at time t. This quantity will be negative if units are backordered, and the vector $X(\cdot) = (X_i(\cdot))$ will be called the inventory process. Given $X_i(0) = C_i$, the inventory process can be expressed as

$$X_i(t) = C_i - Q_i(t) - Z_i(t) , \quad t \geq 0 . \tag{2.3}$$

Let h_i and b_i be the respective costs per unit time of holding and backordering a unit of

product i, and define the cost function

$$H(x) = \sum_{i=1}^{K} H_i(x) \ ,$$

where $H_i(x) = h_i x_i^+ + b_i x_i^-$, $x^+ = \max(x, 0)$ and $x^- = \min(x, 0)$. The production control problem is to choose $C_1, \ldots, C_K$ to minimize the expected average cost

$$J(C_1, \ldots, C_K) = \limsup_{T \to \infty} \frac{1}{T} E \int_0^T H(X(s)) ds \ . \tag{2.4}$$

Since the optimization problem (2.1)-(2.4) does not appear to be tractable without additional distributional assumptions, we turn to the aid of asymptotic analysis and assume that the system operates in heavy traffic; that is, the traffic intensity

$$\rho = \sum_{i=1}^{K} \frac{d_i}{\mu_i} < 1 \tag{2.5}$$

is close to 1. Assumption (2.5) requires that the machine be working the great majority of time to satisfy demand over the long run. Hereafter, the production-distribution system described in this section will be referred to as the *original system S*.

For subsequent use, we introduce the large parameter

$$n_0 = (1 - \rho)^{-2} \tag{2.6}$$

and the rescaled distribution functions of delay times

$$F_i(t) = F_i^0(t(1-\rho)^{-2}) = F_i^0(n_0 t) \ . \tag{2.7}$$

3   A Sequence of Systems in Heavy Traffic

To implement the asymptotic analysis, we introduce a sequence of systems S_n indexed by the parameter n ($n \to \infty$). Each system has the same structure as the original system S. We fix $\{\zeta_j^i\}_{j \geq 1}$, $i = 1, ..., K$, which represent the interdemand intervals, and allow $\{\xi_j^{ni}\}_{j \geq 1}$ $\{\eta_j^{ni}\}_{j \geq 1}$ to depend on n. Thus, the parameters associated with the demand d_i and ν_{di}^2 do not depend on n, while the parameters μ_i^n, ν_{ai}^n and the function $F_i^n(\cdot)$, $i = 1, \ldots, K$ vary with

6

n. To emphasize this dependence we attach an upper index n to the appropriate processes A_i^n, T_i^n etc. The initial inventories C_i^n, $i = 1, \ldots, K$ with respect to which the optimization procedure will be implemented can also depend on n.

Let $\rho_i^n = d_i/\mu_i^n$ and $\rho^n = \sum_{i=1}^K \rho_i^n$. Our main assumption on the sequence S_n are the following:

$$\mu_i^n \to \tilde{\mu}_i, \quad \nu_{ai}^n \to \tilde{\nu}_{ai}, \quad i = 1, \ldots, K, \tag{3.1}$$

$$\eta_j^{ni} = n n_0^{-1} \eta_j^i , \tag{3.2}$$

$$\rho^n < 1 \quad \text{and} \quad \sqrt{n}(1 - \rho^n) \equiv 1 . \tag{3.3}$$

Observe that the distribution function $F_i^n(t)$ of η_1^{ni} is given by $P(\eta_1^{ni} \le t) = F_i(t/n) = F_i^0(n_0 t/n)$, where $F_i(\cdot)$ is defined in (2.7). Assumption (3.3) is the *heavy traffic* condition for the sequence S_n, $n \ge 1$, which requires the traffic intensity ρ^n to converge to 1. As a consequence of (3.1)-(3.3) we have

$$\rho_i^n \to \tilde{\rho}_i \equiv \frac{d_i}{\tilde{\mu}_i} \quad \text{as} \quad n \to \infty, \quad \text{and} \quad \sum_{i=1}^K \tilde{\rho}_i \equiv 1 . \tag{3.4}$$

Assumption (3.2) implies that the delays are increasing as $n \to \infty$.

Let the processes $Q^n(\cdot), Z^n(\cdot), X^n(\cdot)$ associated with system S_n be defined similarly to (2.1)-(2.3):

$$Q_i^n(t) = D_i(t) - A_i^n(T_i^n(t)), \quad Z_i^n(t) = \sum_{j=1}^{A_i^n(T_i^n(t))} I(\eta_j^{ni} + \tau_j^{ni} > t)$$

and

$$X_i^n(t) = C_i^n - Q_i^n(t) - Z_i^n(t) . \tag{3.5}$$

For the processes X^n, Q^n, Z^n and the cost functional associated with the nth system, we will consider two types of rescaling: *Fluid rescaling*

$$\bar{X}^n(t) = \frac{1}{n} X^n(nt) , \bar{Q}^n(t) = \frac{1}{n} Q^n(nt) , \bar{Z}^n(t) = \frac{1}{n} Z^n(nt) , \tag{3.6}$$

$$\bar{J}(C_1^n, \ldots, C_K^n) = n^{-1} \limsup_{T \to \infty} \frac{1}{T} E \int_0^T c(X^n(t)) dt \tag{3.7}$$

7

and *diffusion-type rescaling*

$$\hat{X}^n(t) = \frac{1}{\sqrt{n}} X^n(nt) , \quad \hat{Q}^n(t) = \frac{1}{\sqrt{n}} Q^n(nt) , \quad \hat{Z}^n(t) = \frac{1}{\sqrt{n}} Z^n(nt) , \tag{3.8}$$

$$\hat{J}(C_1^n, \ldots, C_K^n) = n^{-1/2} \limsup_{T \to \infty} \frac{1}{T} E \int_0^T c(X^n(t)) dt . \tag{3.9}$$

In the remainder of this section, we describe the limiting behaviour of the queue length processes Q_i^n associated with product i, $i = 1, \ldots, K$. Since all production orders are served by the machine on a FCFS basis, we can apply existing heavy traffic limit theorems for multiclass systems. Let $I^n(t) = t - \sum_{i=1}^K T_i^n(t)$ denote the cumulative time during $[0, t]$ that the production machine is idle. Then the process defined by

$$W^n(t) = \sum_{i=1}^K \sum_{j=1}^{D_i(t)} \xi_j^{ni} - t + I^n(t) , \quad t \geq 0$$

represents the total amount of work present in the machine's queue at time t. Denote the rescaled version of W^n by $\hat{W}^n(t) = n^{-1/2} W^n(nt)$ and define $\hat{D}_i^n(t) = n^{-1/2}(D_i(nt) - d_i nt)$. For any function $x(\cdot)$ from the space $D[0, \infty)$ let $\Phi(x)$ be the one-dimensional reflection mapping (see Harrison (1985)); that is,

$$\Phi(x)(t) = x(t) + \inf_{s \leq t} [x(s)]^- , \quad t \geq 0 .$$

Define the Brownian motion process

$$W(t) = -t + \bar{W}_i^d(t) + W_i^d(t) , \quad t \geq 0 , \tag{3.10}$$

where $\bar{W}_i^d$ and W_i^d are two independent Wiener processes with variances $\sum_{k=1}^K d_k \tilde{\mu}_k^{-2} \tilde{\nu}_{ak}^2 + \sum_{k=1, k \neq i}^K d_k \tilde{\mu}_k^{-2} \nu_{dk}^2$ and $d_i \tilde{\mu}_i^{-2} \nu_{di}^2$. respectively. Set

$$B(\cdot) = \Phi(W)(\cdot) , \tag{3.11}$$

so that B is a reflected Brownian motion (RBM) on $[0, \infty)$ with drift -1 and variance

$$\sum_{i=1}^K d_i \tilde{\mu}_i^{-2} [\tilde{\nu}_{ai}^2 + \nu_{di}^2] . \tag{3.12}$$

The following result is a consequence of Theorem 1 of Peterson (1991).

8

Proposition 3.1. *Under assumptions (3.1)-(3.3) we have*

$$(\hat{W}^n, \hat{Q}_i^n, \hat{D}_i^n) \xrightarrow{d} (B, d_i B, \tilde{\mu}_i W_i^d) , \quad i = 1, \ldots, K, \quad n \to \infty ,$$

where $\xrightarrow{d}$ stands for weak convergence in $D^3[0, \infty)$.

4 The Limiting Optimization Problem

As mentioned in the Introduction, a rather crude fluid approximation to the control problem is required as a precursor to the heavy traffic diffusion-Gaussian approximation.

4.1. Fluid approximation. For brevity of notation, we introduce $G_i(t) = 1 - F_i(t)$.

Proposition 4.1. *Assume (3.1)-(3.3) and $n^{-1} C_i^n \to c_i^{(1)}$, $i = 1, \ldots, K$. Then for any $C > 0$,*

$$\lim_{n \to \infty} \sup_{t \leq C} |\bar{X}_i^n - x_i(t)| = 0 \quad P - a.s. ,$$

where

$$x_i(t) = c_i^{(1)} - d_i \int_0^t G_i(u) ds . \tag{4.1}$$

Proof. Proposition 3.1 implies that

$$\bar{Q}_i^n(t) \to 0 , \quad n \to \infty , \quad i = 1, \ldots, K \tag{4.2}$$

uniformly on finite time intervals. By the elementary renewal theorem (Karlin and Taylor 1975, page 188), $n^{-1} D_i(nt) \to d_i t$ as $n \to \infty$. Since $\bar{Q}^n(t) = n^{-1}[D_i(nt) - A_i^n(T_i^n(nt))]$, we have

$$n^{-1} A_i^n(T_i^n(nt)) \to d_i t , \quad n \to \infty . \tag{4.3}$$

Equation (3.5) implies

$$\bar{X}_i^n(t) = n^{-1} C_i^n - \bar{Q}_i^n(t) - n^{-1} \sum_{i=1}^{A_i^n(T_i^n(nt))} I(\eta_j^{ni} + \tau_j^{ni} > nt) . \tag{4.4}$$

Observe that $n^{-1} \tau_j^{ni}$, $j \geq 1$ are the jump moments of $A_i^n(T_i^n(n \cdot))$ and $n^{-1} \eta_j^{ni} = n_0^{-1} \eta_j^i$ has the distribution function $F_i^n(nt) = F_i(t)$. Therefore, the last term in (4.4) is the rescaled number of customers in an infinite server queue with service time distribution $F_i(\cdot)$ and

9

arrival stream $A_i^n(T_i^n(nt))$. Limit theorems for such models were considered by Borovkov (1984). Theorem 1 in Chapter 2 of Borovkov implies that

$$n^{-1} \sum_{i=1}^{A_i^n(T_i^n(nt))} I(\eta_j^{ni} + \tau_j^{ni} > nt) \to d_i \int_0^t G_i(t-u)du \equiv d_i \int_0^t G_i(u)du \quad P - \text{a.s.}$$

uniformly on finite time intervals, which together with (4.2) completes the proof. □

Observe that the cost with fluid rescaling can be written as

$$\bar{J}(C_1^n, \ldots, C_K^n) = \limsup_{T \to \infty} \frac{1}{T} E \int_0^T H(\bar{X}^n(t))dt \ .$$

Proposition 4.1 suggests that the following deterministic optimization problem should be considered as the *limiting fluid problem*: Choose $c_1^{(1)}, \ldots, c_K^{(1)}$ to

$$\text{minimize } \limsup_{T \to \infty} \frac{1}{T} E \int_0^T c(x_i(t))dt \ , \tag{4.5}$$

where $x_i(\cdot)$ are given by (4.1). If we set

$$c_i^{(1)} = d_i \int_0^\infty G_i(u)du \ , \tag{4.6}$$

then $\lim_{t \to \infty} x_i(t) = 0$ and the associated cost is equal to zero and is minimal. Since $c_i^{(1)}$ is the limit of $n^{-1} C_i^n$ as $n \to \infty$, we set $C_i^n = n c_i^{(1)} + \Delta C_i^n$, where $\Delta C_i^n = o(n)$, and turn our attention to the next level of approximation.

4.2. Diffusion-Gaussian approximation. For each fixed n, define the processes $Y^n(t) = (Y_1^n(t), \ldots, Y_K^n(t))$

$$Y_i^n(t) = X_i^n(t) - V_i^n(t) \ , \ t \geq 0 \ , \ i = 1, \ldots, K \ , \tag{4.7}$$

where

$$V_i^n(t) = n c_i^{(1)} - n d_i \int_0^{t/n} G_i(u)du$$

and $c_i^{(1)}$ are given by (4.6). Since $V_i^n(t) \to 0$ as $t \to \infty$ for each n, we have

$$\hat{J}^n(C_1^n, \ldots, C_K^n) = n^{-1/2} \limsup_{T \to \infty} \frac{1}{T} E \int_0^T H(X^n(t))dt = n^{-1/2} \limsup_{T \to \infty} \frac{1}{T} E \int_0^T H(Y^n(t))dt \ .$$

$$\tag{4.8}$$

10

Thus, our goal is to minimize the right-hand side of (4.8).

Consider the rescaled process $\hat{Y}^n(t) = n^{-1/2} Y^n(nt)$, $t \geq 0$.

Proposition 4.2 *Assume that $n^{-1/2}\Delta C_i^n \to c_i^{(2)}$, $i = 1, \ldots, K$. Then*

$$\hat{Y}_i^n \xrightarrow{d} \tilde{Y}_i ,$$

where

$$\tilde{Y}_i(t) = c_i^{(2)} - d_i \int_0^t F_i(t-u) dB(u) - \tilde{\mu}_i \int_0^t G_i(t-u) dW_i^d(u) + \theta_i(t) , \qquad (4.9)$$

and $\theta_i(\cdot)$ is a Gaussian process with zero mean and covariance function

$$E\theta_i(t)\theta_i(s) = d_i \int_0^s F_i(s-u) G_i(t-u) du , \quad s \leq t \qquad (4.10)$$

independent of $B(\cdot)$, $W_i^d(\cdot)$ and $\theta_j(\cdot)$, $j \neq i$ $(i,j = 1, \ldots, K)$.

Proof. Definitions (3.5) and (4.7) imply that

$$\hat{Y}_i^n(t) = n^{-1/2}\Delta C_i^n - n^{-1/2} \sum_{i=1}^{A_i^n(T_i^n(nt))} I(\eta_j^{ni} + \tau_j^{ni} > nt) - \hat{Q}_i^n(t) + d_i \int_0^t G_i(t-u) du .$$

Since $n^{-1}\tau_j^{ni}$, $j = 1, 2, \ldots$ are sequential jump moments of the process $A_i^n(T_i^n(n\cdot))$, for any RCLL function $f(\cdot)$ one can write

$$\sum_{j=1}^{A_i^n(T_i^n(nt))} f(n^{-1}\tau_j^{ni}) = \int_0^t f(u) dA_i^n(T_i^n(nu)) ,$$

which is a Lebesgue-Stieltjes integral with respect to a nondecreasing function. Using this fact and (3.2) we derive

$$\hat{Y}_i^n(t) = n^{-1/2}\Delta C_i^n - \int_0^t F_i(t-u) d\hat{Q}_i^n(u) - \int_0^t G_i(t-u) d\hat{D}_i^n(u) + \theta_i^n(t) , \qquad (4.11)$$

where

$$\theta_i^n(t) = n^{-1/2} \sum_{i=1}^{A_i^n(T_i^n(nt))} [I(n_0^{-1}\eta_j^i + n^{-1}\tau_j^{ni} \leq t) - F_i(t - n^{-1}\tau_j^{ni})] . \qquad (4.12)$$

The process θ_i^n can be represented as an integral with respect to a random field, which in this case is a stochastic process with a two-dimensional variable. To derive such a representation, we define the random fields

$$U_i^n(t,x) = n^{-1/2} \sum_{j=1}^{[nt]} [I(F_i(n_0^{-1}\eta_j^i) \leq x) - \max(x,1)] , \quad t \geq 0 , \quad x \geq 0$$

11

and

$$V_i^n(t,x) = n^{-1/2} \sum_{j=1}^{A_i^n(T_i^n(nt))} [I(n_0^{-1}\eta_j^i \le x) - F_i(x)] \ , \ t \ge 0 \ , \ x \ge 0 \ .$$

Hence,

$$V_i^n(t,x) = U_i^n(n^{-1}A_i^n(T_i^n(nt)), F_i(x)) \ . \tag{4.13}$$

Observe that the random variable $F_i(n_0^{-1}\eta_j^i)$ has a uniform distribution on $[0,1]$. In the statistical literature (see Gaenssler and Stute (1979), for example), the process U_i^n is referred to as the stochastic empirical process. The trajectories of $U_i^n(\cdot,\cdot)$ belong to the space $D^2[0,\infty)^2$ (see Pardjanadze and Hmaladze (1986) and Gaenssler and Stute), which is the space of functions $u(t,x)$ defined on $[0,\infty) \times [0,\infty)$ such that the limits

$$u(t+,x+) = \lim_{s\downarrow t, y\downarrow x} u(s,y) \ , \ u(t+,x-) = \lim_{s\downarrow t, y\uparrow x} u(s,y) \ ,$$

$$u(t-,x+) = \lim_{s\uparrow t, y\downarrow x} u(s,y) \ , \ u(t-,x-) = \lim_{s\uparrow t, y\uparrow x} u(s,y)$$

exist and $u(t+,x+) = u(t,x)$. The space $D^2[0,\infty)^2$ endowed with topology similar to Skorohod's J_1 topology for $D[0,\infty)$ is a complete metric space (see Pardjanadze and Hmaladze for details). An increment of a function $u \in D^2[0,\infty)^2$ is defined by

$$\Box u((\tau,y),(t,x)) = u(\tau,y) - u(\tau,x) - u(t,y) + u(t,x) \ .$$

Using this concept one can define the space of functions with bounded variation in the usual way (see Kolmogorov and Fomin (1975)). Observe that the trajectories of the process $(V_i^n(s,x), s \ge 0, x \ge 0)$ belong to $D^2[0,\infty)^2$ and have bounded variation. Therefore the Lebesgue-Stieltjes integral below is well defined and we can write

$$\theta_i^n(t) = \int_0^t I(s+x \le t)dV_i^n(s,t) \ . \tag{4.14}$$

Using the definition of increments given above, one can easily check that (4.14) is equivalent to (4.12). Bickel and Wichura (1971) proved that the empirical process $U_i^n(s,x)$ converges weakly in $D^2[0,\infty)^2$ as $n \to \infty$ to the Gaussian random field $U(s,x)$ with zero mean and covariance function

$$EU(s,x)U(t,y) = s \wedge t(x \wedge y - xy) \ , \ x \le 1, \ y \le 1 \ . \tag{4.15}$$

12

For fixed x, $U(\cdot, x)$ is a standard Brownian Motion, and $U(s, \cdot)$ is a Brownian Bridge for fixed s. Since $U_i^n(\cdot, \cdot)$ is independent of $\hat{Q}_i^n(\cdot)$ and $\hat{D}_i^n(\cdot)$ for each $i = 1, .., K$, Proposition 3.1, (4.3) and (4.13) give the joint convergence

$$(\hat{Q}_i^n(\cdot), \hat{D}_i^n(\cdot), V_i^n(\cdot, \cdot)) \xrightarrow{d} (d_i B(\cdot), \tilde{\mu}_i W_i^d(\cdot), V_i(\cdot, \cdot)) , \quad n \to \infty , \qquad (4.16)$$

where $V_i(t, x) = U(d_i t, F_i(x))$. To complete the proof we need to take the limit as $n \to \infty$ in (4.11). Applying the integration by parts formula and the continuous mapping theorem, one can easily verify that the second and third terms in (4.11) converges to those in (4.9). Define

$$\theta_i(t) = \int_0^t I(s + x \le t) dV_i(s, x) ,$$

where the integral is interpreted as the mean-square integral with respect to a Gaussian field. Then (4.14) and (4.16) imply that

$$\theta_i^n \xrightarrow{d} \theta_i , \quad n \to \infty . \qquad (4.17)$$

Unfortunately, the integration by parts formula cannot be applied directly in this case, since $I(s + x \le t)$ as a function of (s, x) does not belong to $D^2[0, \infty)^2$. Hence, a careful analysis is required to justify this convergence. We do not provide here the proof of this result, since it was carried out in Theorem 2 of Borovkov by different techniques. Readers are referred to Krichagina and Puhalskii (1993) for an alternative proof based on the representation (4.14). The joint convergence of (4.16) and (4.17) allows us to take the limit as $n \to \infty$ in (4.11) and obtain the desired result. It follows from (4.15) that the covariance function $K_i(u, v, x, y)$ of the process $V_i(\cdot, \cdot) = U_i(d_i t, F_i(x))$ is given by

$$K_i(u, v, x, y) = E V_i(u, x) V_i(v, y) = d_i(u \wedge v)(F(x) \wedge F(y) - F(x)F(y)) .$$

The properties of the mean-square integral imply that

$$
\begin{aligned}
E\theta_i(s)\theta_i(t) &= E \int_0^\infty \int_0^\infty I(u + x \le s) dV_i(u, x) \int_0^\infty I(v + y \le t) dV_i(v, y) \\
&= \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty I(u + x \le s) I(v + y \le t) dK_i(u, v, x, y) .
\end{aligned}
$$

13

Since $du \wedge v = I(u = v)du$, $dF_i(x) \wedge F_i(y) = I(x = y)dF_i(x)$ and $dF_i(x)F_i(y) = dF_i(x)dF_i(y)$, the right-hand side for $s \leq t$ is equal to

$$d_i\{\int_0^\infty \int_0^\infty I(u + x \leq s)dudF_i(x) - \int_0^\infty \int_0^\infty \int_0^\infty I(u + x \leq s)I(u + y \leq t)dudF_i(x)dF_i(y)\} .$$

Calculating the integrals we obtain (4.10). $\square$

For each fixed i, let $\bar{F}_i(t)$, $t \leq 0$ be the distribution on $(-\infty, 0]$ generated by $F_i(\cdot)$ as follows:

$$\bar{F}_i(t) = 1 - F_i(-t) \equiv G_i(-t) , \quad t \leq 0 .$$

Denote by γ_i^d and γ_i^{del} (the superscripts represent demand and delay, respectively) two independent Gaussian random variables with zero means and variances $d_i \nu_{di}^2 \int_0^\infty G_i^2(u)du$ and $d_i \int_0^\infty F_i(u)G_i(u)du$, respectively. Let $B^{st}(\cdot)$ be the *stationary* RBM with the same structure and parameters as $B(\cdot)$ (see (3.10)-(3.12)). Since B^{st} is a stationary process we can assume that it is defined for $t \in (-\infty, +\infty)$. Define

$$\gamma_i^B = d_i \int_{-\infty}^0 B^{st}(u)d\bar{F}_i(u) , \tag{4.18}$$

which is the stationary Brownian motion averaged by the distribution function $\bar{F}_i(t) = G_i(-t)$. The following proposition describes the steady-state distribution of $\tilde{Y}_i(\cdot)$.

Proposition 4.3. *Under the assumptions and definitions above, $\tilde{Y}_i(t) \xrightarrow{d} \tilde{Y}_i(\infty)$ as $t \to \infty$, where*

$$\tilde{Y}_i(\infty) = c_i^{(2)} - \gamma_i^B - \gamma_i^d + \gamma_i^{del} . \tag{4.19}$$

Proof. To prove the result we take the limit in (4.9) as $t \to \infty$. Our first step is to show that

$$R_i(t) \equiv \int_0^t F_i(t - u)dB(u) \xrightarrow{d} \int_{-\infty}^0 B^{st}(u)d\bar{F}_i(u) . \tag{4.20}$$

An application of the integration by parts formula yields

$$R_i(t) = \int_0^t F_i(t - u)dB(u) = -\int_0^t B(u)dF_i(t - u) = \int_0^t B(u)dG_i(t - u) . \tag{4.21}$$

For N such that $0 < N \leq t$, define

$$R_i^N(t) = \int_{t-N}^t B(u)dG_i(t - u) . \tag{4.22}$$

14

By Theorem 4.2 of Billingsley, (4.20) will follow from the statements below:

$$R_i^N(t) \xrightarrow{d} \int_{-N}^{0} B^{st}(u)d\bar{F}_i(u) \ , \ t \to +\infty \ \text{ for fixed } N \ ; \tag{4.23}$$

$$\int_{-N}^{0} B^{st}(u)d\bar{F}_i(u) \xrightarrow{d} \int_{-\infty}^{0} B^{st}(u)d\bar{F}_i(u) \ , \ N \to +\infty \ ; \tag{4.24}$$

$$\lim_{N\to\infty} \limsup_{t\to\infty} P(|R_i(t) - R_i^N(t)| > \epsilon) = 0 \ , \ \epsilon > 0 \ . \tag{4.25}$$

To prove (4.23), we start by deriving $R_i^N(t) = \int_0^N B(t - N + u)dG_i(N - u)$ from (4.22). Standard arguments can be used to show that $B(t + \cdot) \xrightarrow{d} B^{st}(\cdot)$ as $t \to \infty$. For fixed N, the continuous mapping theorem gives $R_i^N(t) \xrightarrow{d} \int_0^N B^{st}(u)dG_i(N - u)$, $t \to \infty$. Since B^{st} is a stationary process, we have $\int_0^N B^{st}(u)dG_i(N - u) \stackrel{d}{=} \int_{-N}^0 B^{st}(u)dG_i(-u)$, and (4.23) is proved.

Because $EB^{st}(u) \equiv const$, it follows that $E \int_{-\infty}^{-N} B^{st}(u)d\bar{F}_i(u) = (const)\bar{F}_i(-N) = (const)G_i(N) \to 0$ as $N \to \infty$. Thus, (4.24) is verified.

Equations (4.21) and (4.22) imply that $R_i(t) - R_i^N(t) = \int_0^{t-N} B(u)dG_i(t-u)$. The closed form expression for the distribution of $B(t)$, $t \geq 0$ (see p. 15 of Harrison 1985) implies that $EB(t) \leq const$. Thus $E|R_i(t) - R_i^N(t)| \leq (const)(G_i(N) - G_i(t))$. Applying Chebyshev's inequality and allowing $t \to \infty$ and then $N \to \infty$ yields (4.25), and hence (4.20).

For each $t \geq 0$, the random variables $\tilde{\mu}_i \int_0^t G_i(t - u)dW_i^d(u)$ and $\theta_i(t)$ are Gaussian with zero mean and variances $d_i\nu_{di}^2 \int_0^t G_i^2(u)du$ and $d_i \int_0^t F_i(u)G_i(u)du$, respectively. As $t \to \infty$, they converge to γ_i^d and γ_i^{del}, respectively. The joint convergence follows from the structure of B and the independence of θ_i and B, W_i^d. Thus, one can take the limit in (4.9) as $t \to \infty$ and obtain the result. $\square$

The cost functional given by (4.8) can be expressed in terms of $\hat{Y}^n$ as

$$\limsup_{T\to\infty} \frac{1}{T}E\int_0^T H(\hat{Y}^n(s))ds \ .$$

Based on this fact and Propositions 4.2-4.3, the limiting optimization problem corresponding to the diffusion-Gaussian approximation is to choose $c_1^{(2)}, \ldots, c_K^{(2)}$ to

$$\text{minimize } \limsup_{T\to\infty} \frac{1}{T}E\int_0^T H(\tilde{Y}(s))ds = EH(\tilde{Y}(\infty)) \tag{4.26}$$

15

where $\tilde{Y}(\infty) = (\tilde{Y}_1(\infty), \dots, \tilde{Y}_K(\infty))$ is given by (4.19). The additive structure of the cost and steady-state equalities (4.19) imply that the minimization can be implemented for each component i individually. Thus, (4.26) is equivalent to finding

$$\min_{c_i^{(2)}} \ EH_i(\tilde{Y}_i(\infty)) \tag{4.27}$$

for each $i = 1, \dots, K$. Let $c_i^{(2)}$, $i = 1, \dots, K$ be the solution to (4.27). Then our two-stage asymptotic analysis implies that *nearly optimal* values for the parameters $C_1^n, \dots, C_K^n$ are given by

$$C_i^n = nc_i^{(1)} + \sqrt{n}c^{(2)} \ , \ i = 1, \dots, K \ . \tag{4.28}$$

5 Two-Stage Approximation for the Original System

In this section, we employ the limiting fluid and diffusion-Gaussian problems as approximations for the original optimization problem, and provide numerical solutions under certain additional assumptions. The relation between the original system S and the sequence S_n, $n \geq 1$ is that the dynamics of S can be identified with that of S_{n_0}, where n_0 is given by (2.6), if we set $\mu_i = \mu_i^{n_0}$ and $\nu_{ai}^2 = (\nu_{ai}^{n_0})^2$. Notice, however, that the limiting parameters $\tilde{\mu}_i$ and $\tilde{\nu}_{ai}^2$, $i = 1, \dots, K$ are not directly connected with those of the original system S; they were introduced to justify the limiting procedure in Propositions 3.1 and 4.1-4.2. Therefore, to obtain the *diffusion-Gaussian approximation* for the original system S, we substitute limiting "tilde" parameters by the corresponding parameters of the original system.

The original optimization problem with the cost functional (2.4) normalized by the factor n_0^{-1} can be approximated by the fluid model (4.1), (4.5). As in (4.6), the optimal parameters are

$$c_i^{(1)} = d_i \int_0^\infty G_i(u)du = \frac{d_i}{\lambda_i n_0} \ , \ i = 1, \dots, K \ .$$

In accordance with (4.28), the nearly optimal parameters at the fluid level of approximation are

$$C_i \approx n_0 c_i^{(1)} = \frac{d_i}{\lambda_i} \ . \tag{5.1}$$

16

Since C_i are the initial inventories, this expression admits the following explanation. The system operates in a balanced regime and the average production rate is equal to the average demand rate d_i for each product i. Therefore, the number of units in the delay node plus the number in inventory is approximately equal to the initial inventory C_i for any $i = 1, \ldots, K$. Over the long run, the average number of product i units at the delay node is equal to d_i/λ_i, and therefore product's i's inventory is roughly $C_i - d_i/\lambda_i$. If C_i differs from (5.1), then there would be a constant backlog or a constant excess of inventory over the long run, which would increase the cost functional.

The diffusion-Gaussian approximation corresponds to the original problem with the cost functional (2.4) normalized by the factor $n_0^{-1/2}$. To find the corresponding term $\sqrt{n_0}c_i^{(2)}$ in (4.28), we must solve the minimization problem

$$\min_{c_i^{(2)}} \; EH_i(Y_i(\infty))ds \tag{5.2}$$

for each $i = 1, \ldots, K$, where

$$Y_i(\infty) = c_i^{(2)} - \gamma_i^B - \gamma_i^d + \gamma_i^{del} \;, \tag{5.3}$$

$$\gamma_i^B = d_i \int_{-\infty}^0 B^{st}(u)d\bar{F}_i(u) \;,$$

B^{st} is the stationary RBM on $[0, \infty)$ with drift $-\rho$ and variance $\sum_{k=1}^K d_k \mu_k^{-2}(\nu_{ak}^2 + \nu_{dk}^2)$, and γ_i^d and γ_i^{del} are two independent Gaussian random variables with zero mean and variances $d_i \nu_{di}^2 \int_0^\infty G_i^2(u)du$ and $d_i \int_0^\infty F_i(u)G_i(u)du$, respectively. Once the parameter $c_i^{(2)}$ that minimizes the cost (5.2) is found, we use (4.28) to set

$$C_i = c_i^{(1)}n_0 + c_i^{(2)}\sqrt{n_0} \equiv d_i/\lambda_i + c_i^{(2)}/(1 - \rho) \;. \tag{5.4}$$

The fluid approximation yields the crude deterministic description of the system in terms of the average values, and the diffusion-Gaussian approximation provides the distribution of fluctuations around these average values. Thus, the second term on the right-hand side of (5.3) characterizes the steady-state inventory's fluctuations caused by the variation in the size of the machine's queue. The third term represents the impact of demand variability, and

17

the fourth term quantifies the effect of fluctuations in delay. By (2.7), $F_i(\cdot)$ is the *rescaled* distribution function of the delay times, and it plays an integral role in the definition of these quantities. Due to the rescaling procedure $(t \to nt)$, we observe all processes, as well as the length of delays, in the compressed time scale. Thus, our analysis is oriented towards systems with long and significantly uncertain delays (that is, "wide" distribution functions $F_i^0(\cdot)$).

Unfortunately, problem (5.2) does not admit an explicit solution. The main difficulty is that the random variables γ_i^B and γ_i^d are dependent and their joint distribution is not easily derived. However, the expression for the steady-state distribution (5.3) reveals the structure of the limiting behavior of production-distribution systems with long delays, and can be used for their analysis. To illustrate the use of our analysis, we assume for simplicity that F_i^0 is a uniform distribution on the interval $[(1 - \alpha_i^{del})/\lambda_i, (1 + \alpha_i^{del})/\lambda_i]$, where $\alpha_i^{del} \leq 1$ characterizes the uncertainty of delays. In this case, the rescaled distribution function F_i is uniform on an interval of length $l_i = 2\alpha_i^{del}(1 - \rho)^2/\lambda_i$, and γ_i^B is equal in distribution to

$$\frac{d_i}{l_i} \int_0^{l_i} B^{st}(u)du \equiv \int_0^1 d_i B^{st}(l_i u)du . \tag{5.5}$$

Let $B_i^{st}(\cdot)$ be the *unitless* RBM on $[0, \infty)$ with drift $-\rho$ and variance

$$\Sigma_i^2 = d_i \sum_{j=1}^{K} d_j \mu_j^{-2}(\nu_{ai}^2 + \nu_{di}^2) .$$

Define the unitless constant $L_i = d_i l_i = 2\alpha_i^{del}(1 - \rho)^2(d_i/\lambda_i)$. Since $d_i B^{st}(l_i \cdot) \overset{d}{=} B_i^{st}(L_i \cdot)$, equation (5.5) gives

$$\gamma_i^B \overset{d}{=} \int_0^1 B_i^{st}(L_i u)du = \frac{1}{L_i} \int_0^{L_i} B_i^{st}(u)du . \tag{5.6}$$

The variances of γ_i^d and γ_i^{del} are equal to

$$\nu_{di}^2 L_i/3 \quad \text{and} \quad L_i/6, \tag{5.7}$$

respectively. In the remainder of this paper, we investigate three cases according to the magnitude of L_i.

Case 1. Suppose that the parameter L_i is small. This happens when the delays are not long enough or have low variability, or the demand rate is small. Then the distribution of γ_i^B is

18

close to that of $B_i^{st}(0)$, which is exponential with parameter $2\rho/\Sigma_i^2$. By (5.7), the random variables γ_i^d and γ_i^{del} have small variances and can be neglected. It is easy to calculate that

$$c_i^{(2)} = \frac{\Sigma_i^2}{2\rho} \log(\frac{h_i + b_i}{h_i})$$

minimizes $EH_i(c_i^{(2)} - \gamma_i^B)$. Thus, for small L_i, the variability in the size of the machine's queue has the largest impact on the fluctuations of the inventory.
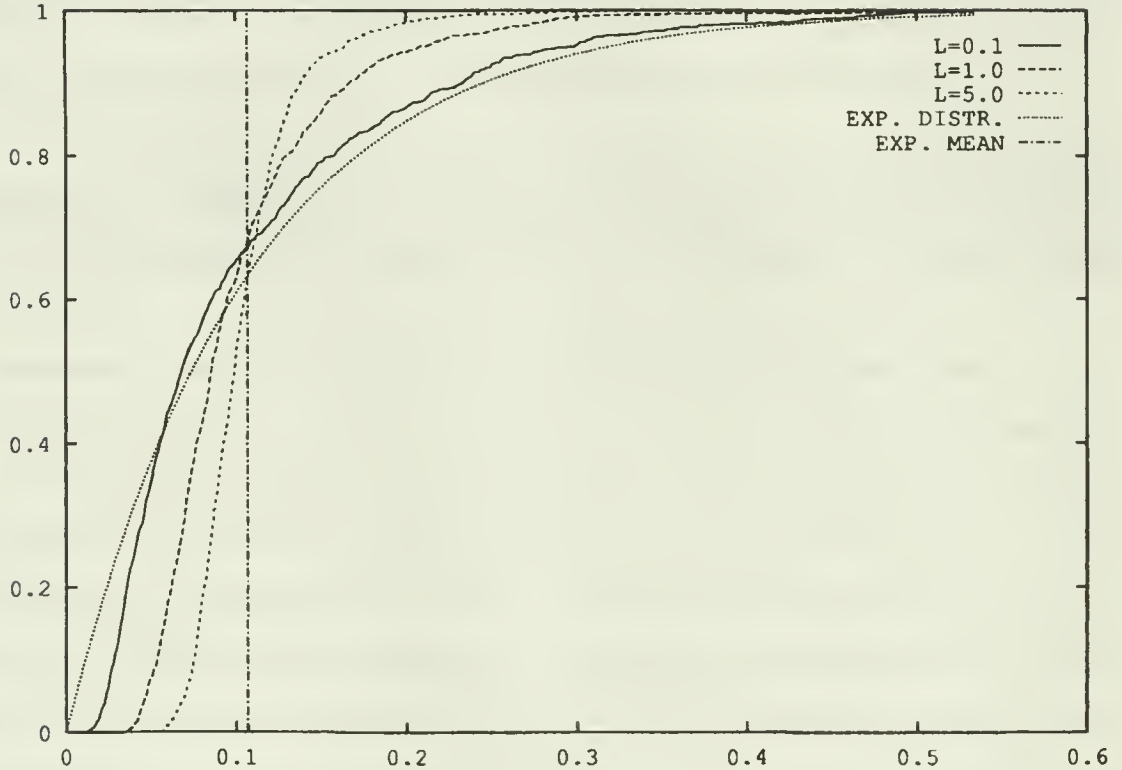

Figure 2. The simulated distributions of γ_i^B for various values of L_i.

Case 2. Suppose that L_i is large. By (5.7), the variances of γ_i^d and γ_i^{del} are large as well. By the Birkhoff-Khinchine theorem, we have

$$\frac{1}{L_i} \int_0^{L_i} B_i^{st}(u)du \xrightarrow{d} EB_i^{st}(\infty) = \frac{\Sigma_i^2}{2\rho} , \quad L_i \to \infty .$$

Hence, by (5.6), the distribution of γ_i^B for large L_i has small variability, and one can let $\gamma_i^B \approx \Sigma_i^2/2\rho$. Since γ_i^d and γ_i^{del} are independent, their sum has a Gaussian distribution with

19

zero mean and variance $L_i(\nu_{di}^2/3 + 1/6)$. The value of the parameter

$$c_i^{(2)} = \text{argmin}_c H_i(c - \frac{\Sigma_i^2}{2\rho} + \gamma_i^d + \gamma_i^{del})$$

is easy to find numerically.

Case 3. Problem (5.2) is more difficult when L_i is neither small nor large. To make further progress in this case, we make the additional assumption that $\gamma_i^d \approx 0$ and can be neglected. That is, ν_{di}^2 is small and the demand is nearly constant. As alluded to earlier, this assumption is required because γ_i^d and γ_i^B are dependent random variables. The distribution of γ_i^B also cannot be calculated explicitly. However, a one-dimensional reflected diffusion is easily simulated.

The simulated distributions of γ_i^B for various values of L_i are presented in Figure 2. As predicted by Case 1, we observe that for small values of L_i the distribution is close to the exponential with parameter $2\rho/\Sigma_i^2$ shown by the dashed line. In accordance with Case 2, Figure 2 also indicates that for larger values of L_i the distribution is concentrated around the mean of the exponential distribution shown by the vertical line. By calculating the cost function $H_i(x)$ averaged by the convolution of the distribution of γ_i^B found by simulation and a normal distribution with mean $c_i^{(2)}$ and variance $L_i/6$, we were able to obtain the cost for any given $c_i^{(2)}$ and find the minimum. Tables 1 and 2 compare our results (COST TH) with the results obtained via simulation of the original system (COST SIM) for one- and two-dimensional problems. The service times were assumed to be uniformly distributed on $[(1 - \alpha_i)/\mu_i, (1 + \alpha_i)/\mu_i]$, $i = 1, \ldots, K$. In all four problems, our cost minimizing values of C_i are very close to those obtained via simulation.

Problem 1 parameters: $\mu_1 = 0.9^{-1}, d_1 = 1.0, \alpha_1 = 0.8, \alpha^{del} = 0.8, 1/\lambda_1 = 50.0$.
Number of simulated jobs : 2.5×10^5.

C_1	50.0	51.0	52.0	53.0	54.0	55.0	56.0	57.0	58.0	59.0
COST SIM	10.41	8.36	6.94	6.09	5.77	5.89	6.32	7.00	7.80	8.70
COST TH	11.15	8.79	7.09	6.03	5.52	5.48	5.80	6.38	7.15	8.03

Problem 2 parameters: $d_1 = 1.0, 1/\lambda = 100.0, \mu_1 = 0.9^{-1}, \alpha = 0.8, \alpha^{del} = 0.5$.
Number of simulated jobs: 5.0×10^5.

C_1	100.0	101.0	102.0	103.0	104.0	105.0	106.0	107.0	108.0	109.0	110.0
COST SIM	11.29	9.30	7.84	6.96	6.57	6.59	6.93	7.50	8.23	9.07	9.88
COST TH	12.09	9.76	8.04	6.88	6.24	6.04	6.22	6.67	7.33	8.13	9.03

Problem 3 parameters: $d_1 = 1.0, 1/\lambda_1 = 100.0, \mu_1 = 0.9^{-1}, \alpha = 0.8, \alpha^{del} = 0.99$.
Number of simulated jobs: 7.5×10^5.

C_1	100.0	101.0	102.0	103.0	104.0	105.0	106.0	107.0	108.0	109.0	110.0
COST SIM	14.94	12.97	11.41	10.25	9.46	9.01	8.86	8.98	9.31	9.82	10.47
COST TH	15.8	13.59	11.79	10.40	9.41	8.78	8.48	8.46	8.69	9.14	9.70

Table 1: Three one-dimensional problems.

Parameters: $\mu_1 = 1.0$, $\mu_2 = 0.8$, $d_1 = 0.5$, $d_2 = 0.32$, $(\rho = 0.9)$, $1/\lambda_1 = 1/\lambda_2 = 50.0$, $\alpha_1 = 0.6$, $\alpha_2 = 0.8$, $\alpha_1^{del} = \alpha_2^{del} = 0.8$.

Number of simulated jobs: 2.0×10^5.

COST SIM:

C_2 $\quad$ C_1	21.0	26.0	27.0	28.0	29.0	30.0	35.0
11.0	53.20	34.36	32.85	32.17	32.17	32.64	37.24
16.0	31.47	11.00	11.12	10.44	10.44	10.90	15.49
17.0	29.13	10.47	8.96	8.28	8.28	8.76	13.36
18.0	28.27	9.43	7.91	7.24	7.24	7.71	12.31
19.0	28.14	9.30	7.78	7.11	7.11	7.58	12.18
20.0	28.62	9.78	8.27	7.59	7.59	8.06	12.67
25.0	33.35	14.51	13.00	12.32	12.32	12.79	17.40

COST TH:

C_2 $\quad$ C_1	21.0	26.0	27.0	28.0	29.0	30.0	35.0
11.0	48.47	31.31	30.32	30.00	30.40	31.06	35.89
16.0	27.60	10.44	9.45	9.21	9.52	10.19	15.02
17.0	25.88	8.72	7.73	7.50	7.80	8.74	13.30
18.0	25.25	8.09	7.10	6.86	7.17	7.84	12.67
19.0	25.41	8.25	7.26	7.02	7.33	8.00	12.83
20.0	26.06	8.90	7.91	7.60	7.98	8.65	13.48
25.0	30.93	13.77	12.78	12.60	12.85	13.52	18.30

Table 2: A two-dimensional problem.

Acknowledgment

References

[1] Bickel P. and M. Wichura. 1971. Convergence Criteria for Mutiparametric Stochastic Processes and Some Applications. *Annals of Math. Statistics* **42**, 1285-1295.

[2] Billingsley, P. 1968. *Convergence of Probability Measures*. John Wiley and Sons, New York.

[3] Borovkov, A. A. 1984. *Asymptotic Methods in Queueing Theory*. John Wiley and Sons, New York.

[4] Clarke, A. J. and H. Scarf. 1960. Optimal Policies for a Multi-Echelon Inventory Problem. *Management Science* **6**, 475-490.

[5] Gaenssler, P. and Stute W. 1979. Empirical processes: a survey of results for independent and identically distributed random variables. *Annals of Probability* **7**, 193-243.

[6] Harrison, J. M. 1985. *Brownian Motion and Stochastic Flow Systems*. John Wiley and Sons, New York.

[7] Harrison, J. M. 1988. Brownian Models of Queueing Networks with Heterogeneous Customer Populations, in W. Fleming and P. L. Lions (eds.), *Stochastic Differential Systems, Stochastic Control Theory and Applications* IMA Volume **10**, Springer-Verlag, New York, 147-186.

[8] Karlin, S. and H. M. Taylor. 1975. *A First Course in Stochastic Processes*. Second Edition, Academic Press, New York.

[9] Kelly. F. P. 1979. *Reversibility and Stochastic Networks*. John Wiley and Sons, New York.

[10] Kelly F. P. and C. N. Laws. 1993. Dynamic Routing in Open Queueing Networks: Brownian Models, Cut Constraints and Resource Pooling. *Queueing Systems* **13**, 47-86.

[11] Kolmogorov, A. N. and Fomin S. V. 1975. *Introduction to Real Analysis*. Dover Publ. Inc., New York.

[12] Krichagina, E. V. and A. A. Puhalskii. 1993. Fluid and Gaussian Approximations for a Queueing System with G/∞ Service Center. Preprint. Institute of Control Sciences, Moscow, Russia.

[13] Krichagina, E. V., R. Rubio, M. I. Taksar and L. M. Wein. 1992. A Dynamic Stochastic Stock Cutting Problem. Sloan School of Management, MIT, Cambridge, MA.

[14] Pardjanadze, A. M. and Hmaladze A. V. 1986. On the Asymptotic Theory of a Statistics of Sequential Ranks. *Theor. Prob. Appl.* **31**, 758-772.

[15] Peña, A. and P. Zipkin. 1992. Dynamic Scheduling Rules for a Multiproduct Make-to-Stock Queue. Graduate School of Business, Columbia University, New York.

[16] Peterson, W. P. 1991. A Heavy Traffic Limit Theorem for Networks of Queues with Multiple Customer Types. *Mathematics of Operations Research* **16**, 90-118.

[17] Veatch, M. H. and L. M. Wein 1993. Scheduling a Make-to-Stock Queue: Index Policies and Hedging Points. To appear in *Operations Research*.

[18] Wein, L. M. 1992a. Scheduling Networks of Queues: Heavy Traffic Analysis of a Multistation Network with Controllable Inputs. *Operations Research* **40**, S312-S334.

[19] Wein, L. M. 1992b. Dynamic Scheduling of a Multiclass Make-to-Stock Queue. *Operations Research* **40**, 724-735.

[20] Zheng, Y. and P. Zipkin. 1990. A Queueing Model to Analyze the Value of Centralized Inventory Information. *Operations Research* **38**, 296-307.